A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit

Anna S. Law, PhD[1], Yvonne Freer, RN PhD[2], Jim Hunter, PhD[3],

Robert H. Logie, PhD[1], Neil McIntosh, DSc(Med)[2], John Quinn, BA(Hons)[2]

From: [1]Department of Psychology, University of Edinburgh,
[2]Department of Neonatology, Royal Infirmary of Edinburgh,
[3]Department of Computing Science, University of Aberdeen

Address correspondence to:
Professor Robert Logie,
Department of Psychology,
University of Edinburgh,
7 George Square,
Edinburgh, EH8 9JZ, U.K.
Email: rlogie@staffmail.ed.ac.uk
Tel: 0131 6511394

Accepted for publication in the Journal of Clinical Monitoring and Computing

ABSTRACT

Objective: To compare expert-generated textual summaries of physiological data with trend graphs, in terms of their ability to support neonatal Intensive Care Unit (ICU) staff in making decisions when presented with medical scenarios.

Methods: Forty neonatal ICU staff were recruited for the experiment, eight from each of five groups – junior, intermediate and senior nurses, junior and senior doctors. The participants were presented with medical scenarios on a computer screen, and asked to choose from a list of 18 possible actions those they thought were appropriate. Half of the scenarios were presented as trend graphs, while the other half were presented as passages of text. The textual summaries had been generated by two human experts and were intended to describe the physiological state of the patient over a short period of time (around 40 minutes) but not to interpret it.

Results: In terms of the content of responses there was a clear advantage for the Text condition, with participants tending to choose more of the appropriate actions when the information was presented as text rather than as graphs. In terms of the speed of response there was no difference between the Graphs and Text conditions. There was no significant difference between the staff groups in terms of speed or content of responses. In contrast to the objective measures of performance, the majority of participants reported a subjective preference for the Graphs condition.

Conclusions: In this experimental task, participants performed better when presented with a textual summary of the medical scenario than when it was presented as a set of trend graphs. If the necessary algorithms could be developed that would allow computers automatically to generate descriptive summaries of physiological data, this could potentially be a useful feature of decision support tools in the intensive care unit.

Key words: intensive care, computerised monitoring, decision making, decision support

INTRODUCTION

Medical and nursing staff working in an Intensive Care Unit (ICU) have access to a large volume of physiological information for the patients in their care. Computerized patient monitoring systems allow this information to be updated as often as every second while displaying cumulative historical data patterns. Across different ICUs, a variety of different computerised monitoring systems have been set up to manage these data [1, 2], but a common approach is that the separate physiological parameters (e.g., heart rate, core and peripheral temperatures, transcutaneous oxygen and carbon dioxide, blood pressures) are displayed on a bedside monitor as a set of trend graphs (e.g., [3]). While in the past the different parameters might have been displayed on separate monitors, most computer monitoring systems can display all of the data on a single monitor for a range of physiological functions, and with the different parameters displayed in comparable formats. The use of a single, integrated display was found to be one of the perceived benefits of computerised systems in a survey of their use in ICUs across a range of European countries [4]. Another benefit is that trend monitoring systems can store data from any given patient over his or her entire stay in intensive care. The data collected automatically from the patient can be supplemented with manual entry of notes from staff regarding medical history, aetiology, treatment and care regimes along with demographic, diagnostic and prognostic details. The clinician therefore potentially has access to a full set of patient details, historical and current, when reaching decisions, and may be able to interact with the system to enter further comments on the patient's condition, make a note of tests to be conducted, or add test results.

The continued development of this technology is intended to support ICU staff in decisions regarding patient care. However, past research has shown that the introduction of systems that display data trends does not necessarily lead to clinical improvements [5, 6]. Cunningham et al. [5] conducted a clinical trial where 600 babies admitted to a neonatal ICU were randomly allocated to one of four groups. For one group there was no display of trend data, for another group there was a continuous display of trend data, and for the other two groups there was an alternation between trend data and no trend data every 24 hours. None of outcome measures showed any advantage to the patients of computerized physiological trend monitoring during their stay in intensive care, even when the outcome after 1 to 4 years was examined. The main perceived benefits of the trend monitoring system were as an aid to research and to staff education.

McIntosh, Lyon and Badger [6] suggested that data overload may be one of the factors that undermines the usefulness of trend monitoring in the neonatal ICU. Alberdi et al. [7] noted that different grades of staff may respond to a trend monitoring system in very different ways, and staff grade may predict how efficiently graphical monitoring systems are used. The participants in Alberdi et al.'s study were shown trend graphs from a period of two hours and asked to "think aloud" while viewing them. Junior and senior doctors were equally likely to identify "key" events in the physiological traces, but the senior doctors were more likely to identify subtle events that were nevertheless relevant. When staff were observed on the ward, it was the senior doctors who referred to the monitoring system most often and who were most knowledgeable about how to use it. Nurses and junior doctors spent more time on the ward than senior doctors, but seemed to benefit less from the trend monitoring system, and consulted it rarely. All grades of staff had difficulty identifying the onset of adverse trends as they were developing, but could identify when a trend had commenced when looking at them in retrospect.

Ewing et al. [8] and Freer et al. [9] have shown that different grades of staff categorize and think about information in crucially different ways, and also use different kinds of information in fulfilling their duties. Nurses focus on patient information that cannot readily

be monitored automatically, such as sleep/wake states, muscle tone and movement, respiratory effort, or skin pallor. They are also involved in the administration of care that might appear as artefacts in the trend monitoring (e.g. changing probes), and that adds to the difficulty of interpreting the trends displayed. Nurses also appear to view the computerized monitoring system primarily as a tool for senior doctors (see also [3, 7]). This emphasizes the need to question whether trend graphs of selected physiological functions are the most appropriate format for use on the ward.

The experiment reported here investigated whether textual summarisation might be an alternative means of presenting patient information, that could better facilitate the task of interpreting the state of the patient and deciding upon the actions that need to be taken. The summaries in question were generated by human experts, because the algorithms have yet to be fully developed that would allow a computer to produce a natural language summary of complex time series data. The textual summaries were then compared with trend graphs in an "off-ward" experiment where doctors and nurses had to say what actions they would take based on the information presented in a medical scenario. Participants completed half of the scenarios with the graphical presentation and half with the textual presentation. If the Text condition were found to produce superior performance, this could be taken as encouragement that the possibility of automated natural language summarisation of ICU data is worth pursuing further. Also, it would lead us to consider whether the graphical displays that dominate trend monitoring systems could usefully be supplemented, or even in some cases replaced, by textual summaries. Textual summaries are also easier and less expensive than graphical displays to transmit to remote sites - for example to a senior doctor who might be at another hospital. On the other hand, if a textual summary created by a human expert was less useful or no more useful than the graphs, then this might offer evidence for the utility of graphical displays in this context rather than simply assuming that such displays are best because the technology makes them available.

Five staff groups were recruited for the experiment reported here – junior nurses, intermediate nurses, senior nurses, junior doctors and senior doctors. As the medical scenarios in the experiment were based around actions that would normally be taken by both nurses and doctors, no prediction was made about group differences in overall success. However, it was possible that doctors and nurses would be more likely to identify the need to take an action when it was normally part of their own job. For example, a nurse might be more likely than a doctor to notice that the baby was cold, and that the incubator temperature should be increased. The major purpose of the experiment was in any case to explore the possibility that a different pattern of performance would be obtained with textual presentation of the medical scenarios than with graphical presentation. The staff were familiar with the use of a trend monitoring facility known as the BADGER system, and displays were constructed and presented in a research version of this system known as the Time Series Workbench [10].

METHOD

Participants

Participants were recruited on a voluntary basis from among staff working at the neonatal ICU at the Royal Infirmary of Edinburgh, Scotland, where the current research took place. There were eight participants from each of five groups, namely Junior Nurses (JN), Intermediate Nurses (IN), Senior Nurses (SN), Junior Doctors (JD) and Senior Doctors (SD). Nurses were classified as junior, intermediate or senior on the basis of their number of years of neonatal experience. Those with three or fewer years experience were classified as *junior*, those with between four and fourteen years experience were classified as *intermediate* and those with fifteen or more years experience were classified as *senior*. *Junior* doctors in this experiment were Senior House Officers, with one years neonatal experience or less, while the neonatal experience of the *senior* doctors ranged from 4 years to 25 years. All the nurses were female except for two of the junior group. Six of the eight junior doctors were female, as were three of the senior doctors. Two other participants (one junior nurse, one senior nurse) completed one session of the experiment, but were not able to return for the second session, and replacements were recruited.

Scenarios

Sixteen medical scenarios were chosen to represent two examples of each of the eight "main target" actions, namely: order chest X-ray, intubate or re-intubate, re-apply transcutaneous probe, start dopamine, treat with surfactant, put baby on High Frequency Oscillatory Ventilation (HFOV), start Continuous Positive Airway Pressure (CPAP), or No Action. The physiological data for the scenarios had previously been recorded from babies who had been cared for in the unit, and there was also a record of the actions that had originally been taken on the ward for those patients. This had been collected by a research nurse who observed the care of these patients and noted any actions taken or observations made by the staff [11,12]. The (anonymised) data for all 16 scenarios were already available for display as trend graphs showing heart rate, transcutaneous oxygen and carbon dioxide, oxygen saturation, core and peripheral temperatures and mean blood pressure. To create the textual condition, two of us (NM, a consultant neonatologist and YF, an experienced neonatal nurse and clinical researcher), produced a descriptive summary of the graphs in each scenario. This summary was designed to be descriptive only, so that participants still had to do the work of interpreting the physiological data and deciding what action (if any) was required. In generating the text, the following assumptions were made:

- participants know reference ranges for all measurements, including blood gas results;
- participants can tell which data points are artifactual; this includes probe recalibrations and blood gas sampling as well as dropouts;
- participants are familiar with the shape of bradycardia and desaturation etc on the traces, and can distinguish them from artifact; for example, they can distinguish motion artifact on the pulse oximeter from a genuine desaturation;
- participants can tell the severity of these events i.e. from looking at the chart they are able to tell whether a bradycardia is mild or severe.

Any statements which were couched in these terms were considered to be descriptive (even though to a naïve reader they might seem to contain an element of interpretation). As a check we had all textual summaries segmented and the segments evaluated as to the degree of interpretation contained in them by a semi-independent rater. This rater (author JQ), was a computer scientist involved in developing software for neonatal ICU monitoring, and who was therefore familiar with the environment and the forms of data being presented, but did not have any clinical experience or any other involvement in the experiment. He judged that 357

of the 373 segments were purely descriptive (under the above assumptions). Most of the comments that he picked out as being interpretative made a link between an action and its effect on the data. Where the relationship was "obvious" (such as changes in FiO2 setting and oxygen saturation) these were not considered to be interpretative.

The time-period covered by the scenarios varied between 30 minutes and 53 minutes with an average of 40.5 (sd = 6.5) minutes. Participants were asked to say what actions should be taken at the end of the time period depicted or described. One scenario in each condition was designed to provoke the response of "No Action", as the patient was stable. In addition to the description of the physiological data for the designated time-period, "background" information was also provided for the patient in the scenario. This background was presented as text, and contained information such as weeks of gestation, days since birth, weight, and recent actions taken on the ward. The background summary was exactly the same for both the Graphs and Text conditions, and was shown in a panel on the left hand side of screen. An example scenario is shown in Figure 1, in both its graphical and textual form. Half of the participants in the experiment saw scenarios 1-8 in graphical form and scenarios 9-16 in textual form, while the opposite was the case for the other half of the participants. The order in which the Graphs and Text conditions were attempted was also counterbalanced.

Insert Figure 1 about here

In response to the medical scenarios, participants chose from a multiple-choice list of 18 possible actions, displayed at the bottom of the screen (see Table 1). This list was chosen from a lexicon of 50 actions (established as discussed in Ewing et al. [8]), as clearly the entire lexicon would have been unmanageable in the context of a multiple-choice response situation. The actions were the same for every scenario, and were always presented in alphabetical order. Although the scenarios were chosen to represent particular types of actions, there was usually more than one action from the list that was appropriate for the scenario. For example, scenario 1 was chosen as an example of an occasion where the main target action was *order a chest X-ray* (this was what had happened on the ward originally). However, it was also appropriate to say that the baby needed to be warmed, as there was a wide gap between core and peripheral temperatures. Therefore, in addition to the main target action, NM and YF provided for each scenario, an analysis of other "appropriate" actions; these lists were taken as the "gold standard" with which participants' performance was compared. NM and YF also determined which of these actions were considered to be part of the primary responsibilities of a nurse or of a doctor, or were equally relevant for both. This allowed us to analyse whether all the staff groups were equally likely to be able to identify certain actions that were a) part of a nurse's job or b) part of a doctor's job. Table 1 shows the 18 actions and whether these were "nurse" or "doctor" actions (or "both").

Insert Table 1 about here

For every scenario participants were given a maximum time of three minutes to respond. This time limit was introduced not to impose time pressure, but in order to guarantee the maximum length of an experimental session. It was important to reassure nurses and doctors coming away from the ward that the experimental session would only take half an hour. Pilot work was conducted prior to the experiment to determine that three minutes was an adequate length of time for participants to complete each scenario. When responding, participants were first asked to choose the actions that they would take themselves ("What would you do?"), and

then asked to say what actions they thought others would take ("What would you expect others to do?"). So for Scenario 1, nurses might choose "warm baby" for themselves, and "CXR" for the "others", as ordering a chest X-ray is the responsibility of a doctor. Participants had to click on a button that read "Accept" after each stage of responding, and complete both stages within the three minutes. This two-stage responding process was introduced in order to establish each participant's complete understanding of the situation, not just their understanding of their own responsibilities. However, it became clear that participants often accidentally chose an action during the "wrong" stage. In particular, the doctors tended to forget to press the Accept button for their own actions before choosing the "nurse" actions. This may be due to the tendency for nurse actions to be considered first in reality, for example if a baby is in respiratory distress the nurse might initiate handbagging while the doctor is summoned to intubate the baby. Also, it was often the case that a participant would notice something else that they should have put in under the first stage of responding, and select it while in the second stage. Therefore in the analysis reported below, no account is taken of the stage at which the participant selected a particular action – their responses during the three minutes of the scenario are considered as a whole.

Equipment/software

The computer used to present the medical scenarios was a standard Dell PC running Windows XP professional, with a 17 inch monitor, at a resolution of 1024 x 768 pixels. The scenarios were presented using a software tool called the Time Series Workbench (TSW) [10] (see Figure 1). The TSW also recorded which actions participants chose, and the times at which they chose them. A feature of the TSW display is that the user could click on one of the physiological traces causing a pop-up box to appear with the exact value. Beneath the graphs are coloured markers indicating events that occurred on the ward when the babies were originally observed. The user could also click on these markers causing a box to pop up saying, for example "incubator open" or giving the result of a test. These were the only types of interaction that the participant could have with the display, other than choosing their responses.

Procedure

The experiment had a mixed design, with two levels of the within-subjects factor of presentation condition (i.e., Graphs vs. Text) and five levels of the between-subjects factor of staff group. The experiment was conducted in a meeting room within the neonatal unit. Participants attended individually, and attempted each condition of the experiment in separate sessions, completed on different days. In general the time gap between the two conditions ranged from one day to three weeks, but one junior doctor did attempt both conditions in the same day, and one of the senior doctors had a gap of 31 days between conditions. A one-way ANOVA showed that there was no significant difference between the groups in terms of the time gap between the two conditions, $F(4, 35) = 1.08$, ns, $MSE = 51.636$. All sessions took around half an hour each and the order in which participants attempted the presentation conditions was counterbalanced. At the beginning of each session participants were shown a general instruction screen, which explained that they would be shown physiological information for a series of babies, and that they had to decide on appropriate actions based on this information. The next part of the session was an exercise to familiarise them with the list of actions from which they would choose for each medical scenario. The complete list was displayed at the bottom of the screen in the same (alphabetical) order in which it would appear throughout the experiment. In the centre of the screen the actions from the list appeared one at a time in a random order. As each action appeared, participants had to locate that action in the list at the bottom of the screen, and click on the check box beside it using the mouse. This exercise was intended to reduce some of the variance in response times by familiarising the participants with the items on the list, and with their locations. Participants

went through the list twice in their first session with the actions shown one at a time in the centre of the screen in a different random order on the second presentation. In the second session, they went through the list only once.

Following the familiarisation exercise, participants were given the instructions for the medical scenarios. The instructions varied according to the condition (Graphs or Text) that participants were attempting at the time. They were instructed that they would see the graphs or text on the right hand side of the screen, and a background summary about the baby on the left. They were told that the baby might be improving, deteriorating or stable, and that they could choose as many actions as they wanted from the bottom of the screen, or select "No Action" if they felt this was appropriate. After selecting the appropriate actions at both the "You" and "Others" stages of responding, they were asked to click on the "Accept" button. They were warned that they could not de-select a choice once it had been made. Participants were given two practice scenarios in each condition, presented as either graphs or as text, depending on the condition. The practice scenarios were the same across sessions but were different from those used for the main part of the experiment. They then completed the eight scenarios for the condition that they were attempting. At the end of the second session they were asked to report their subjective preference between the graphical and textual presentation.

RESULTS

Speed of Responses

Time Outs

The TSW software recorded the timing of all responses to the nearest hundredth of a second. If a participant had not pressed the Accept button for the second time at three minutes (to indicate that their response was complete), the scenario "timed out". The number of time-outs for each staff group in each condition is shown in Table 2. Twenty out of the forty participants were not timed out on any of their sixteen scenarios. The maximum number of time-outs observed was ten out of a possible sixteen, but this occurred for only one participant, while the remainder had four or fewer time-outs. In almost all of these cases, the participant made some response before three minutes; it was very rare for a scenario to time out with no response at all. In fact this only happened on four occasions, two with the same participant and all with graphical presentation of the scenario. Due to the positive skew in the data set, a non-parametric sign test was used to examine whether time-outs were more common in one condition of the experiment. This showed that there was no significant difference between the number of people who were timed out more often in the Graphs condition ($N = 10$) and the number who were timed out more often in the Text condition ($N = 5$).

It was also possible that the staff grade might influence the likelihood of the scenario timing out. The junior nurses had the highest median number of time-outs at six, while the intermediate nurses and junior doctors had a median of three, and the senior nurse and senior doctor groups each had a median of four. A Kruskal-Wallis ANOVA showed that there was no significant difference in the median number of time-outs for each staff group, $H(4, N = 40) = 5.016$, $p = 0.285$.

Time to completion

The 46 scenarios where the participant had timed out were removed from the data set of 640 scenarios. The time-to-completion was then examined – the mean for the Graphs condition

was 113.76 seconds (sd = 23.85) and the overall mean for the Text condition was almost exactly the same at 113.18 seconds (sd = 20.58); the maximum possible time was 180 seconds. The timing data split up by staff group are shown in Table 2 – the junior and senior nurse groups had faster mean reaction times than the other groups. A 2x5 mixed ANOVA (where the within-subjects factor was Graphs or Text condition and the between subjects factor was the five staff groups) was conducted to determine whether these group differences were significant. This showed that there was no main effect of group, $F(4, 35) = 1.907$, ns, MSE = 922.421, no main effect of condition, $F(1, 35) = 0.462$, ns, MSE = 196.121 and no interaction, $F(4, 35) = 0.258$, ns, MSE = 196.121. There was no tendency for either the presentation condition or the staff group to influence the speed with which the scenario was completed.

Insert Table 2 about here

Content of Responses

Main target actions

The scenarios were originally chosen as leading to particular main target actions:(1) order chest X-ray, (2) intubate or re-intubate, (3) re-apply transcutaneous probe, (4) start dopamine, (5) treat with surfactant, (6) put baby on HFOV, (7) start CPAP or (8) No Action. There were two scenarios leading to each of the eight actions. Therefore the first analysis of content was a simple count of how many of these "main target" actions were identified by participants in each condition. The maximum possible score was 8, but this was only achieved by one participant, a senior doctor in the Text condition. The data for each group are displayed in Table 2. It is clear that more main target actions were identified in the Text condition than the Graphs condition – the overall mean for the Graphs condition is 3.13 actions (sd = 1.18) and the overall mean for the Text condition is 4.80 (sd = 1.52). The junior doctor group has the smallest gap between performance in the two conditions, as can be seen in Figure 2.

Insert Figure 2 about here

A 2x5 mixed ANOVA was conducted with two levels of the within-subjects factor of Condition (Graphs vs. Text) and 5 levels of the between-subjects factor Group (junior, intermediate and senior nurses, junior and senior doctors). There was a highly significant main effect of condition, $F(1,35) = 32.835$, $p < 0.001$, MSE = 1.709. There was no main effect of group, $F(4, 35) = 1.296$, ns, MSE = 1.702. The interaction between group and condition only approached significance, $F(4, 35) = 2.425$, $p = 0.066$, MSE = 1.709. So, there was no significant difference between the groups, and overall participants performed better in the Text condition than the Graphs condition.

Total proportion of appropriate actions identified

The next dependent measure was of the proportion of all the "appropriate" actions that participants chose in each scenario. So for example, the appropriate actions for Scenario 1 are "warm baby" and "order chest X-ray", according to our experts. Therefore if a participant only identified "warm baby", he or she would score 0.5 for that scenario against a maximum possible of 1.0. A participant's scores were averaged across the eight scenarios in each condition. Where the appropriate action was "No Action", then a choice of "observe" and/or "minimal handling" was also accepted as an appropriate answer. The proportion data for each group are shown in Table 2. The overall mean for the Graphs condition was 0.37 (sd = 0.13)

and the overall mean for the Text condition was 0.57 (sd = 0.12). There is a clear advantage for the Text condition. A 2x5 mixed ANOVA was conducted which showed a significant main effect of presentation condition, $F(1, 35) = 30.979$, $p < 0.001$, MSE = 0.020, but no significant effect of staff group, $F(4, 25) = 0.016$, ns, MSE = 0.022, and no interaction, $F(4,35) = 1.236$, ns, MSE = 0.020. Therefore, there was no difference between the different staff groups, and they all performed better with the textual presentation than the graphical presentation.

Proportion of appropriate "nurse" and "doctor" actions identified

The next measure was the proportion of appropriate "nurse actions" that were selected by the participants for each scenario. For example, in Scenario 6 the appropriate actions were "warm baby", "recalibrate BP dome" and "put baby on HFOV". The first two of these are generally the responsibility of a nurse, while a doctor would generally take the decision to put the baby on High Frequency Oscillatory Ventilation. The questions of interest in this analysis were whether doctors were just as likely as nurses to identify "nurse" actions, and whether one presentation condition was more likely than the other to lead people to identify this type of action. So, if a participant (doctor or nurse) selected "warm baby" (but not "recalibrate BP dome") in scenario 6 then he or she would score 0.5. These proportion scores were then averaged across the eight scenarios in each condition. Overall, participants scored 0.33 on average in the Graphs condition (sd = 0.18) and 0.55 in the Text condition (sd = 0.16). The proportion scores for each staff group are shown in Table 2. The Text condition performance is clearly better for all groups apart from the junior doctors, who only perform slightly better with the Text condition. It also appears as though the doctor groups are worse overall in identifying nurse actions. A 2x5 ANOVA showed that there was a significant main effect of condition, $F (1, 35) = 47.373$, $p < 0.001$, MSE = 0.024, but no significant main effect of Group, $F (4, 35) = 1.061$, ns, MSE = 0.036, and no interaction, $F (4, 35) = 0.605$, ns, MSE = 0.024. Therefore, the doctor groups were not significantly worse than the nurse groups at selecting nurse actions, although their mean performance is lower with both textual and graphical presentation.

The proportion of appropriate "doctor" actions identified was also examined. For example, in Scenario 6 the appropriate action that would generally be part of a doctor's job was "put baby on HFOV". Therefore participants (doctors and nurses) scored 1 if they identified this action and zero if they did not. The proportion scores were then averaged across the eight scenarios in each condition. The overall mean for the Graphs condition was 0.42 (sd = 0.15), and the overall mean for the Text condition was 0.54 (sd = 0.21). The mean proportion for each group is shown in Table 2. The pattern is slightly different in form to that obtained in the previous analyses, as the junior doctors actually perform better in the Graphs condition than the Text condition. However, a 2x5 ANOVA showed that there was once again a main effect of condition $F (1, 35) = 8.667$, $p = 0.006$, MSE = 0.035, but no main effect of group $F (4, 35) = 0.899$, ns, MSE = 0.032, and no significant interaction, $F (4,35) = 2.107$, $p = 0.101$, MSE = 0.035.

Total number of actions chosen

Using the scoring system of the "proportion of appropriate actions chosen", means that no account is taken of any irrelevant actions that were chosen. In theory a person could achieve a perfect score by selecting every single action for every single scenario. Although there was no sign of such a drastic strategy, it was important to know whether a higher number of actions were chosen overall in the Text condition. The data in Table 2 suggest that this is indeed the case, and a 2x5 ANOVA confirmed that there was a significant main effect of presentation condition $F(1, 35) = 11.344$, $p < 0.002$, MSE = 17.77, but no significant effect of staff group, $F (4, 35) = 1.214$, ns, MSE = 56.71, and no interaction, $F(4, 35) = 0.0307$, ns, MSE = 56.71.

In the Text condition the average number of actions selected was 26.03 (sd = 6.27) and in the Graphs condition the average number selected was 22.85 (sd = 5.93).

Proportion of chosen actions that were appropriate

Given that significantly more actions were chosen for the Text condition than for the Graphs, the better results for the Text condition that have been reported in previous sections could have arisen because participants simply chose more actions in response to the text and scored more appropriate ones by chance. The data were therefore re-analysed according to the proportion of the actions chosen that were appropriate. The drawback with analysing the data in this way is that a person can get a perfect score even if they had an incomplete understanding of the scenario. For example, Scenario 1 has two appropriate actions "warm baby" and "order CXR". If the only action chosen was "warm baby", the participant would score 1. Participants essentially had a "free pass" to choose "minimal handling" or "observe" in this analysis – these responses were not counted among the total because clearly it would never be "inappropriate" to observe a baby in an ICU with minimal handling.

If the Text condition produced better performance with this scoring system as well as with scoring by the proportion of appropriate actions identified (as above), this would be good evidence that the text really was helping participants understand the scenarios and find the optimal solutions. The overall mean for the Graphs condition 0.38 (sd = 0.14), while the overall mean for the Text condition was 0.51 (sd = 0.14). The data for each group are shown in Table 2. A 2x5 mixed ANOVA showed a significant main effect of condition, $F(1, 35) = 15.663$, $p < 0.001$, MSE = 0.021. There was no significant main effect of group, $F(1, 35) = 0.526$, ns, MSE = 0.019 and no interaction, $F(1, 35) = 0.544$, ns, MSE = 0.021. This is the same pattern of results as when the data were analysed by the proportion of the appropriate actions that were chosen. It seems that the Text condition was genuinely helping people to find the appropriate actions, rather than just leading them to choose actions indiscriminately.

Reported preference

Although the Text condition clearly elicited better performance than the Graphs condition, it was also of interest to know whether this was the impression that the participants themselves had of the experiment. All participants were asked at the end of the second condition that they attempted, which presentation format they preferred. Table 3 shows that participants in all staff groups tended to prefer the Graphs condition, despite performing better with the text.

Insert Table 3 about here

DISCUSSION

The experiment showed a clear advantage for the Text condition, whether the data were analysed in terms of the proportion of the appropriate actions that participants identified, or the proportion of the actions they identified that were appropriate. This advantage for the Text condition was found in spite of the fact that all groups reported a preference for the Graphs condition. Participants chose more actions overall when using the textual display, but it seems that this was because it was helping them to identify the actions that, in the opinion of our experts, were appropriate for that scenario. Given the limited list of possible actions, it may have been that participants were sometimes of the opinion that an action not provided on the list was the most appropriate for a particular scenario – a few of them did report that this was a source of frustration. However the "main target" action for each scenario was not only appropriate in the opinion of our own experts, but also in the opinion of the practitioners who

had originally cared for the babies whose data we used in the experiment - the scenarios had been chosen because they presented a time leading up to that action being taken on the ward.

Although performance in the Text condition was superior in terms of the content of responses, there was no difference between graphs and text in terms of the speed of responses – the two conditions produced surprisingly similar response times. It is also perhaps surprising that there was no significant difference between the staff groups on any of the measures. There was a tendency for the junior doctors to get the least benefit from the Text condition, as they had the smallest gap between graphs and text performance for the "main target" actions identified and for the total proportion of appropriate actions identified. In terms of the "doctor actions" identified, the juniors actually performed better with the graphs than the text, although the interaction term was not significant. The results do seem somewhat different from those of Alberdi et al. [7], who found that senior doctors were more likely than juniors to be able to identify all relevant physiological events from graphical information. Here, the junior doctors did not significantly differ from the senior doctors, and if anything, had a higher mean performance in the Graphs condition than the latter, whichever dependent measure is considered.

In an experiment comparing graphical presentation with tabular presentation in a managerial decision task [13], it appeared that graphs were better in terms of conveying overall relationships, but that tables were better if exact values were required. The Text condition in our experiment gave exact values for the physiological data at time points where our experts had deemed it appropriate. However, that information was also contained within the graphical presentation – participants merely had to click on the graph at the time they were interested in for the value to appear. This was carefully explained and demonstrated during the instructional phase of each experimental session. Nevertheless, it did require a bit of extra effort for the participant to retrieve this information. One possible reason why graphical presentation did not produce such good performance could be that participants were not able to easily see from the scale of the graph that a particular physiological parameter had reached a critical point. Scaling is an important issue for graphical presentation and can make a difference to how information is perceived [6, 14]. In the BADGER system used on the ward, users can adjust the scaling on the graphs. However, this might require more time and effort than reading a value reported by a textual summarising system. McIntosh, Lyon and Badger [6] argued that "the more data which are displayed on the screen, the more confusing the screen becomes, particularly to the new system user". The textual summarisation picked out the most relevant information for each of the physiological parameters; there was therefore less data for participants to integrate into their overall understanding of the scenario.

Hanson and Marshall [2] review decision support systems that have been developed within the field of artificial intelligence, and acknowledge that the medical profession has been very wary of adopting systems that seem to take the practice of medicine out of the hands of a human expert and entrust it to a computer. However, they argue that data-driven decision support tools (that use the available data to develop rules and solutions, rather than having these pre-programmed) can be used not as a replacement for a human expert but as an "intelligent assistant". This type of decision support tool could allow doctors and nurses to make full use of the information available to them, thereby improving patient care. The data reported here show that doctors and nurses in a neonatal ICU were better able to find the relevant information among the textual summary than the graphical display. This is an encouraging sign that, assuming a decision support tool could be developed to produce natural language summaries of complex graphical data, it might be of benefit to ICU staff. Clearly, such an approach would have to take into account what is already known about the way people understand language. Wright, Jansen and Wyatt [14] have described some of the ambiguities and errors that can arise in the interpretation of textual data, such as the use of vague quantifiers or the tendency for positive phrases to be easier to understand than negative phrases.

Finally we must address the question as to whether, in comparing graphical and textual presentations of the same data, we are comparing "like with like"? The answer clearly is ***no***, in that the text was generated (with considerable effort) by human experts working from the graphical presentations. The amount of transformation from graph to text is considerable. However, as discussed earlier, we have tried to make sure that this transformation consisted solely of filtering and summarising the information (i.e. reducing its volume by several orders of magnitude) followed by text generation; clearly filtering and summarising require expertise, but we tried as far as possible to make sure that this expertise did not manifest itself in the text as additional interpretation. If the text always had to be generated by hand, it would not be possible to exploit the benefits of textual presentation, as a human expert would always have to be on hand to generate it, and might as well make the decision! However we have reason to believe that automatic text generation of this quality is possible [15]. In this case, the advantages of this format could be achieved at little additional cost.

ACKNOWLEDGEMENTS

This work was part of the NEONATE project which is funded through the People at the Centre of Communication and Information Technologies (PACCIT) program, by both the UK Economic and Social Research Council (ESRC) and the Engineering and Physical Sciences Research Council (EPSRC). Author JQ is supported by the Morton-Stewart Fund. We would very much like to thank the medical and nursing staff of the Neonatal Unit at the Royal Infirmary of Edinburgh for their participation in this research.

REFERENCES

1. Green CA, Gilhooly KJ, Logie RH, Ross DG. Human factors and computerisation in Intensive Care Units: a review. Int J Clin Monit Comput 1991; 8: 167-178

2. Hanson CW, Marshall BE. Artificial intelligence applications in the intensive care unit, Crit Care Med 2001; 29: 427-435

3. Alberdi E, Gilhooly K, Hunter J, Logie R, Lyon A, McIntosh N et al. Computerisation and decision making in neonatal intensive care: A cognitive engineering investigation. J Clin Mont Comput 2000; 16: 85-94

4. Ambroso C, Bowes C, Chambrin MC, Gilhooly K, Green C, Kari A et al. INFORM: European survey of computers in Intensive Care Units. Int J Clin Monit Comput 1992; 9: 53-61

5. Cunningham S, Deere S, Symon A, Elton RA, McIntosh N. A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. Crit Care Med 1998; 26: 2053-2059

6. McIntosh N, Lyon A, Badger P. Time trend monitoring in the Neonatal Intensive Care Unit: Why doesn't it make a difference? Pediatrics 1996; 98:540

7. Alberdi E, Becher JC, Gilhooly K, Hunter J, Logie RH, Lyon A et al. Expertise and the interpretation of computerized physiological data: implications for the design of computerized monitoring in neonatal intensive care. Int J Hum-Comput St 2001; 55: 191-216

8. Ewing G, Freer Y, Logie RH, Hunter J, McIntosh N, Rudkin S et al. Role and Experience Determine Decision Support Interface Requirements in a Neonatal Intensive Care Environment. J Biomed Inform 2003; 36: 240-249

9. Freer Y, Ferguson L, Ewing G, Hunter J, Logie RH, Rudkin S et al. Mismatched concepts in a neonatal intensive care unit (NICU): Further issues for computer decision support? J Clin Monit Comput 2003; 17: 441-447

10. Hunter J. The Time Series Workbench: User Manual, University of Aberdeen Computing Science Technical Report, 2004

11. Ewing G, Ferguson L, Freer Y, Hunter J, McIntosh N. Observational data acquired on a Neonatal Intensive Care Unit. University of Aberdeen Computing Science Departmental Technical Report: TR 0205, 2002

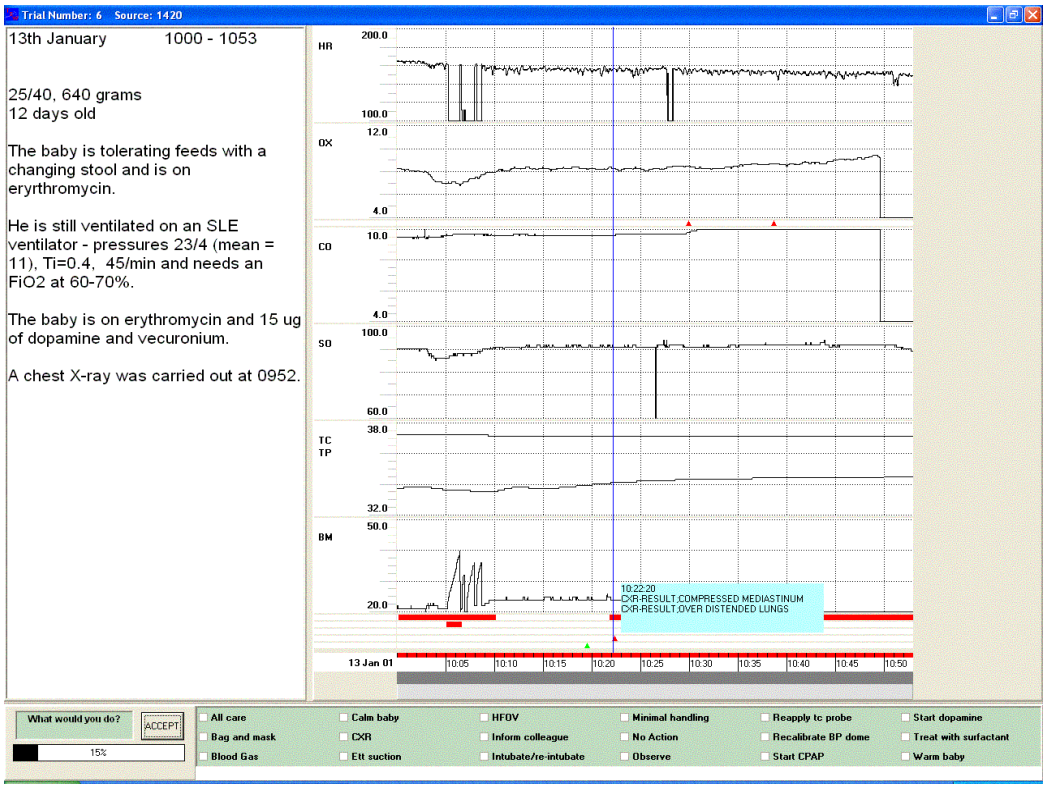
12. Hunter JRW, Ferguson L, Freer Y, Ewing G, Logie R, McCue P et al. The NEONATE Database. Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care, AIME-03, Cyprus, 2003, pp 21-24

13. Benbasat I, Dexter AS, Todd P. The influence of color and graphical information presentation in a managerial decision simulation. Human-Computer Interaction 1986; 2: 65-92

14. Wright P, Jansen C, Wyatt JC. How to limit clinical errors in interpretation of data. The Lancet 1998; 352: 1539-1543

15. Somayajulu G, Sripada S, Reiter E, Hunter J, Yu J. Summarizing neonatal time series data. In: Proceedings of the research note sessions of the EACL03, Budapest, 2003, pp. 167-170

Figure 1: Screenshots of graphical and textual presentation of a scenario



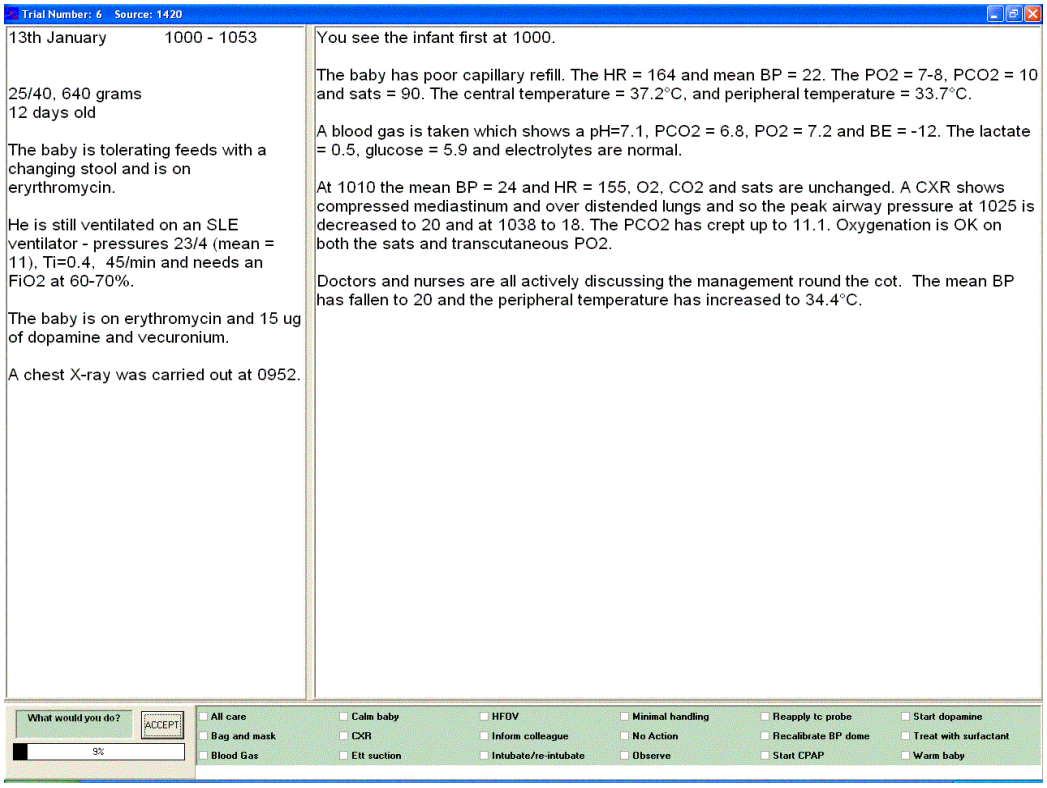

Trial Number: 6 Source: 1420
13th January 1000 - 1053
25/40, 640 grams
12 days old
The baby is tolerating feeds with a changing stool and is on eryrthromycin.
He is still ventilated on an SLE ventilator - pressures 23/4 (mean = 11), Ti=0.4, 45/min and needs an FiO2 at 60-70%.
The baby is on erythromycin and 15 ug of dopamine and vecuronium.
A chest X-ray was carried out at 0952.
You see the infant first at 1000.
The baby has poor capillary refill. The HR = 164 and mean BP = 22. The PO2 = 7-8, PCO2 = 10 and sats = 90. The central temperature = 37.2°C, and peripheral temperature = 33.7°C.
A blood gas is taken which shows a pH=7.1, PCO2 = 6.8, PO2 = 7.2 and BE = -12. The lactate = 0.5, glucose = 5.9 and electrolytes are normal.
At 1010 the mean BP = 24 and HR = 155, O2, CO2 and sats are unchanged. A CXR shows compressed mediastinum and over distended lungs and so the peak airway pressure at 1025 is decreased to 20 and at 1038 to 18. The PCO2 has crept up to 11.1. Oxygenation is OK on both the sats and transcutaneous PO2.
Doctors and nurses are all actively discussing the management round the cot. The mean BP has fallen to 20 and the peripheral temperature has increased to 34.4°C.
What would you do?
ACCEPT
9%
All care
Bag and mask
Blood Gas
Calm baby
CXR
Ett suction
HFOV
Inform colleague
Intubate/re-intubate
Minimal handling
No Action
Observe
Reapply tc probe
Recalibrate BP dome
Start CPAP
Start dopamine
Treat with surfactant
Warm baby

Figure 2: Mean number of "main target" actions identified by each group in both Graphs and Text conditions

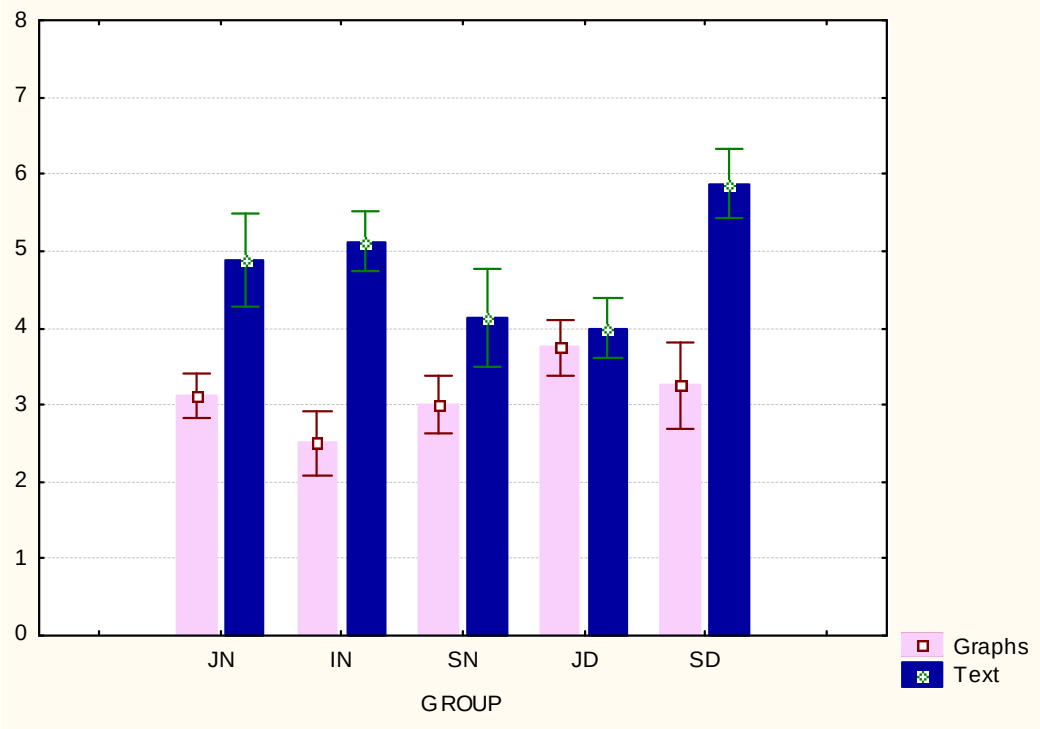


Table 1: List of 18 possible actions

Actions	**Explanation**	**Classified as "nurse", "doctor" or both**
All care	Involves attending to basic hygiene needs & position change	Nurse
Bag and Mask	Resuscitation and stabilisation using hand ventilation using a face mask	Both
Blood Gas	Sampling of blood from an indwelling catheter	Both
Calm baby	Techniques using containment and stroking	Nurse
CXR	Ordering an X-ray of the chest	Doctor*
ETT suction	Aspiration of secretions from an endo-tracheal tube	Nurse
HFOV	High Frequency Oscillatory Ventilation	Doctor*
Inform colleague	Communication with a colleague	Both
Intubate/ reintubate	Inserting an endo-tracheal tube into the trachea	Doctor*
Minimal Handling	Intervening with the infant and or his environment the least possible times	Both
No Action	No action taken	Both
Observe	Baby is being observed but no other action is being taken	Both
Re-apply tc probe	Re-apply transcutaneous probe	Nurse
Recalibrate BP dome	Recalibrate Blood Pressure Dome	Nurse
Start CPAP	Continuous positive airway pressure ventilation	Both
Start dopamine	Intravenous dopamine started	Doctor*
Treat with surfactant	Artificial surfactant instilled into the endo-tracheal tube	Doctor*
Warm baby	Use of artificial means to increase a baby's temperature	Nurse

* These actions can also be taken by an Advanced Neonatal Nurse Practitioner

Table 2: Table of means for all groups on key dependent measures

	Junior Nurses		Intermediate Nurses		Senior Nurses		Junior Doctors		Senior Doctors		All Groups	
	Graphs	Text	Graphs	Text	Graphs	Text	Graphs	Text	Graphs	Text	Graphs	Text
Time to completion (seconds)	**106.14** (25.33)	**106.30** (16.61)	**123.08** (24.30)	**119.38** (14.33)	**102.42** (19.81)	**99.54** (17.50)	**120.33** (30.42)	**117.23** (29.75)	**116.81** (15.74)	**123.45** (15.36)	**113.76** (23.85)	**113.18** (20.58)
Number of time-outs (out of a possible 8)	**0.13** (0.35)	**0.13** (0.35)	**0.63** (0.74)	**0.50** (0.76)	**0.50** (0.53)	**0.13** (0.35)	**1.25** (1.83)	**1.38** (1.85)	**0.88** (0.99)	**0.00** 0.00	**0.68** (1.05)	**0.43** (1.01)
Main Target Actions (out of a possible 8)	**3.13** (0.83)	**4.88** (1.73)	**2.50** (1.20)	**5.13** (1.13)	**3.00** (1.07)	**4.13** (1.81)	**3.75** (1.04)	**4.00** (1.07)	**3.25** (1.58)	**5.88** (1.25)	**3.13** (1.18)	**4.80** (1.52)
Proportion of Appropriate Actions	**0.36** (0.08)	**0.59** (0.12)	**0.35** (0.13)	**0.60** (0.13)	**0.42** (0.13)	**0.55** (0.12)	**0.40** (0.17)	**0.50** (0.11)	**0.33** (0.15)	**0.62** (0.12)	**0.37** (0.13)	**0.57** (0.12)
Proportion of Nurse Actions	**0.34** (0.15)	**0.57** (0.12)	**0.35** (0.22)	**0.66** (0.16)	**0.40** (0.20)	**0.57** (0.11)	**0.33** (0.19)	**0.49** (0.16)	**0.23** (0.14)	**0.52** (0.23)	**0.33** (0.18)	**0.55** (0.16)
Proportion of Doctor Actions	**0.38** (0.12)	**0.58** (0.24)	**0.35** (0.15)	**0.57** (0.19)	**0.41** (0.17)	**0.49** (0.26)	**0.53** (0.11)	**0.42** (0.19)	**0.45** (0.20)	**0.66** (0.15)	**0.42** (0.16)	**0.54** (0.21)
Total Number of Actions Chosen (across 8 scenarios)	**25.75** (4.10)	**29.38** (5.20)	**22.38** (3.42)	**25.38** (5.21)	**23.25** (5.23)	**25.00** (5.13)	**23.50** (8.40)	**26.13** (7.51)	**19.38** (6.72)	**24.25** (7.42)	**22.85** (5.93)	**26.03** (6.27)
Proportion of Actions that were Appropriate	**0.36** (0.14)	**0.46** (0.10)	**0.37** (0.12)	**0.54** (0.13)	**0.42** (0.12)	**0.53** (0.17)	**0.41** (0.15)	**0.47** (0.11)	**0.37** (0.12)	**0.56** (0.17)	**0.38** (0.14)	**0.51** (0.14)

Mean in bold, standard deviation in parenthesis

Table 3: Reported preferences for Graphs or Text

Preference	**Graphs**	**Text**
Junior Nurses	6	2
Intermediate Nurses	5	3
Senior Nurses	5	3
Junior Doctors	6	2
Senior Doctors	7	1
Total	**29**	**11**